Analysis of the ToothGrowth data set

Porter

March 5, 2016

Load Dataset and Basic Summary

First we need to make sure that the ToothGrowth dataset is available. The code below loads the dataset library and loads the ToothGrowth dataset. We then show a basic summary of the dataset.

```
library(datasets)
data("ToothGrowth")
summary(ToothGrowth)
```

```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

Lets take a look at the head of this dataset so we have a good idea of what the entries look like.

```
head(ToothGrowth)
```

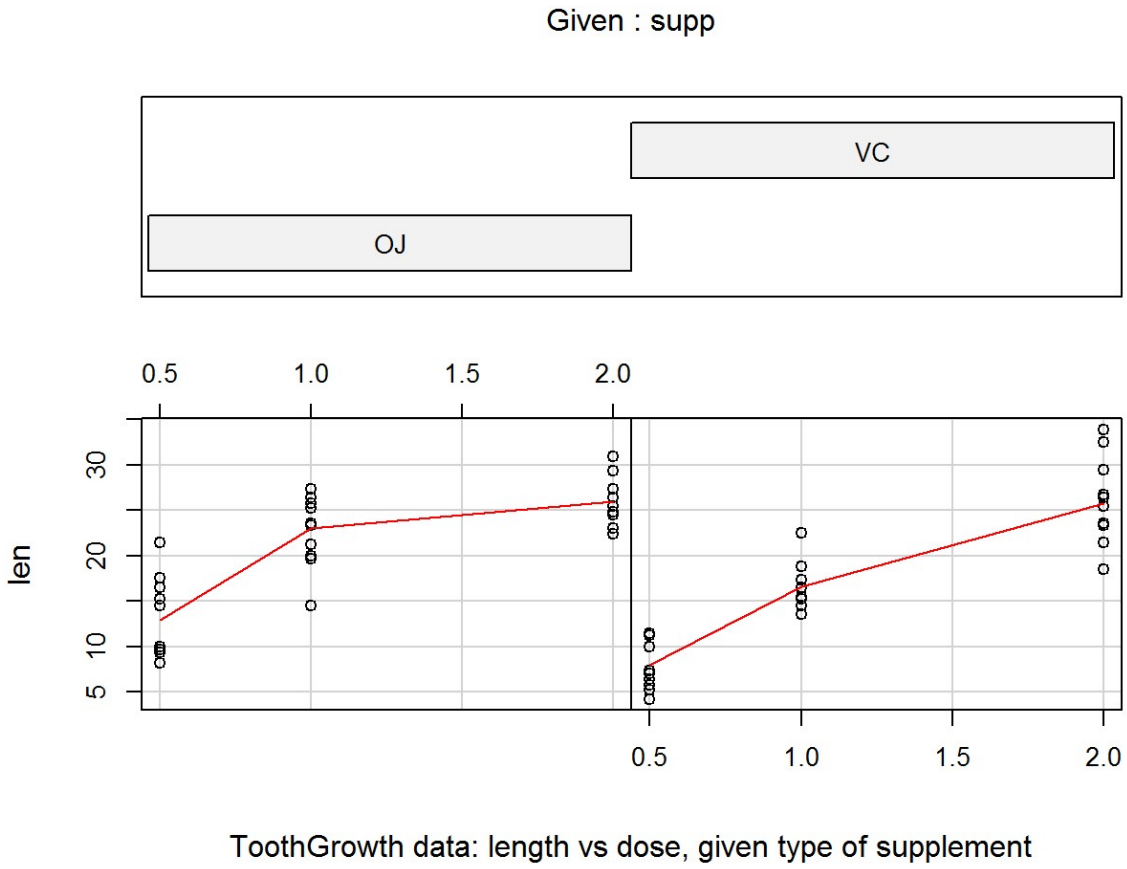
```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
str(ToothGrowth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

Now we can see there are is a ‘len’ variable which is numeric. There is a ‘supp’ variable which is the supplement type and it is a factor with two levels, VC or OJ. Then there is a ‘dose’ variable which is also a numeric. Looking online there is a little more description about what these values are. The ‘len’ variable is the tooth length. The ‘supp’ variable is which supplement type and the two possibilities are VC which is ascorbic acid or vitamin C and OJ which orange juice. The following plot shows the len vs dose with respect to the type of supplement.

```
coplot(len ~ dose | supp, data = ToothGrowth, panel = panel.smooth, xlab = "Too
thGrowth data: length vs dose, given type of supplement")
```



Supplement and Dose comparison tests

Now we want to compare the length given a supplement type and dose. There are a lot of different combinations we could compare. For each supplement there are 3 different dosages, 0.5, 1 and 2. The above coplot shows that in each supplement group the higher the dosage the higher the average tooth length so we are mainly going to compare the two different groups at the equal dosages.

Dosage of 2 mg/day

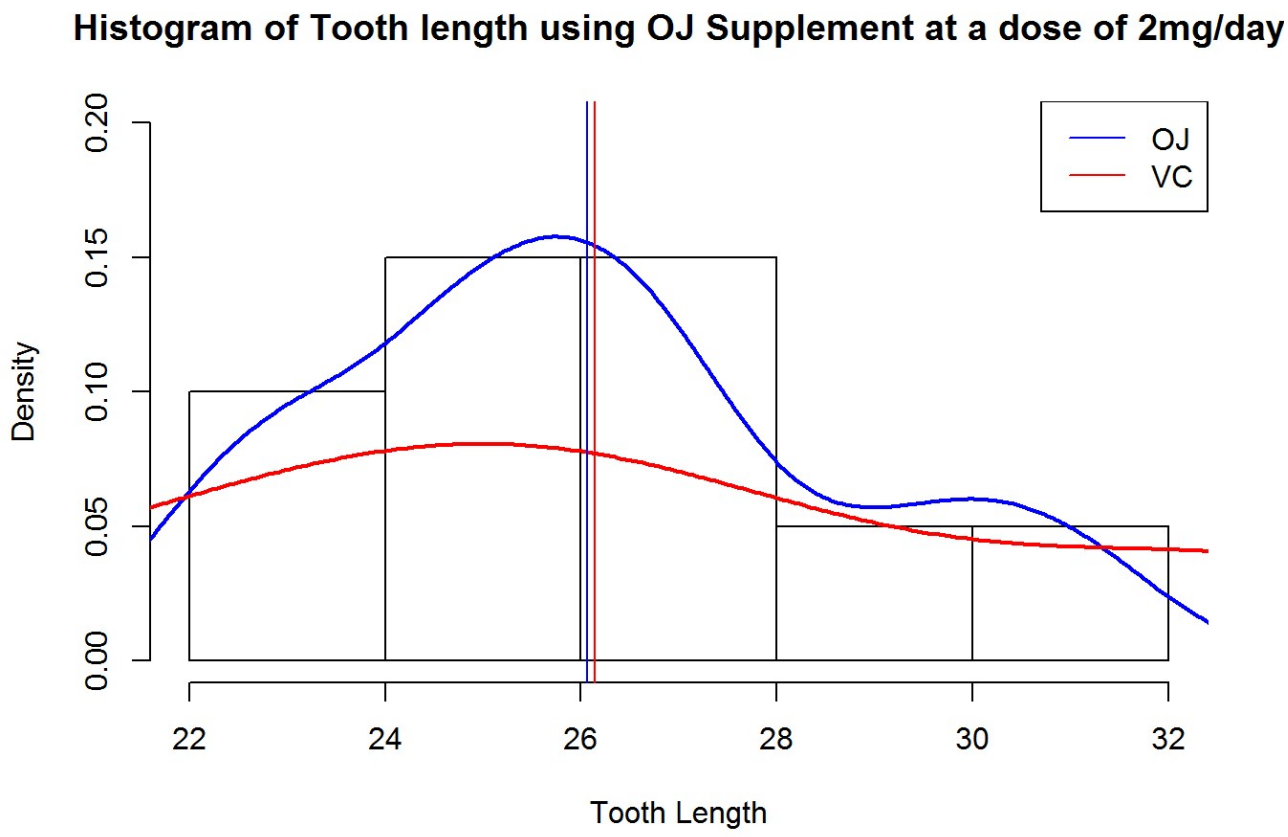
The first thing we will do is subset the dataset so we have just the dosage of 2. The reason we do this is that the we want the t.test to compare equal doses instead of comparing across all doses.

```
toothDose2 <- subset(ToothGrowth, dose == 2)
dose2test <- t.test(len ~ supp, paired = F, var.equal=F, data = toothDose2)
print(dose2test)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

This hypothesis test has the null hypothesis that the difference in the means is 0 and the alternative that the difference is not 0. This test essential tries to answer the question is there a statistically significant difference in the tooth length if one supplement is take versus the other. There are a few things we can look at to see if we should accept or reject the null hypothesis. Let's start by looking at the t-statistic which is -0.0461361. In this test the t-statistic is the estimate of the difference in the average tooth length between the two supplements at dose 2 minus the hypothsized difference which is zero in this test. This value is quite low so we can start to think that we should accept the null hypothesis. The other thing we can look at is the p-value reported for this test. In this case the p-value is representative of the probability of seeing this t-statistic under the null hypothesis. The p-value for this test was 0.9638516 which is high. Meaning that it is very likely that we see the t-statistic under the null hypothesis. The code below will visualize this data. It shows the two distributions and their means.

```
toothDose2_vc = toothDose2[toothDose2$supp == 'VC',]
toothDose2_oj = toothDose2[toothDose2$supp == 'OJ',]
hist(toothDose2_oj$len, probability = T, ylim=c(0,0.2),
     main='Histogram of Tooth length using OJ Supplement at a dose of 2mg/da
y', xlab = "Tooth Length")
lines(density(toothDose2_oj$len), col='blue', lwd=2)
lines(density(toothDose2_vc$len), col='red', lwd=2)
abline(v = dose2test$estimate[1], col='blue')
abline(v = dose2test$estimate[2], col='red')
legend('topright', c('OJ', 'VC'), lty = 1, col=c('blue', 'red'))
```



Dosage of 1 mg/day

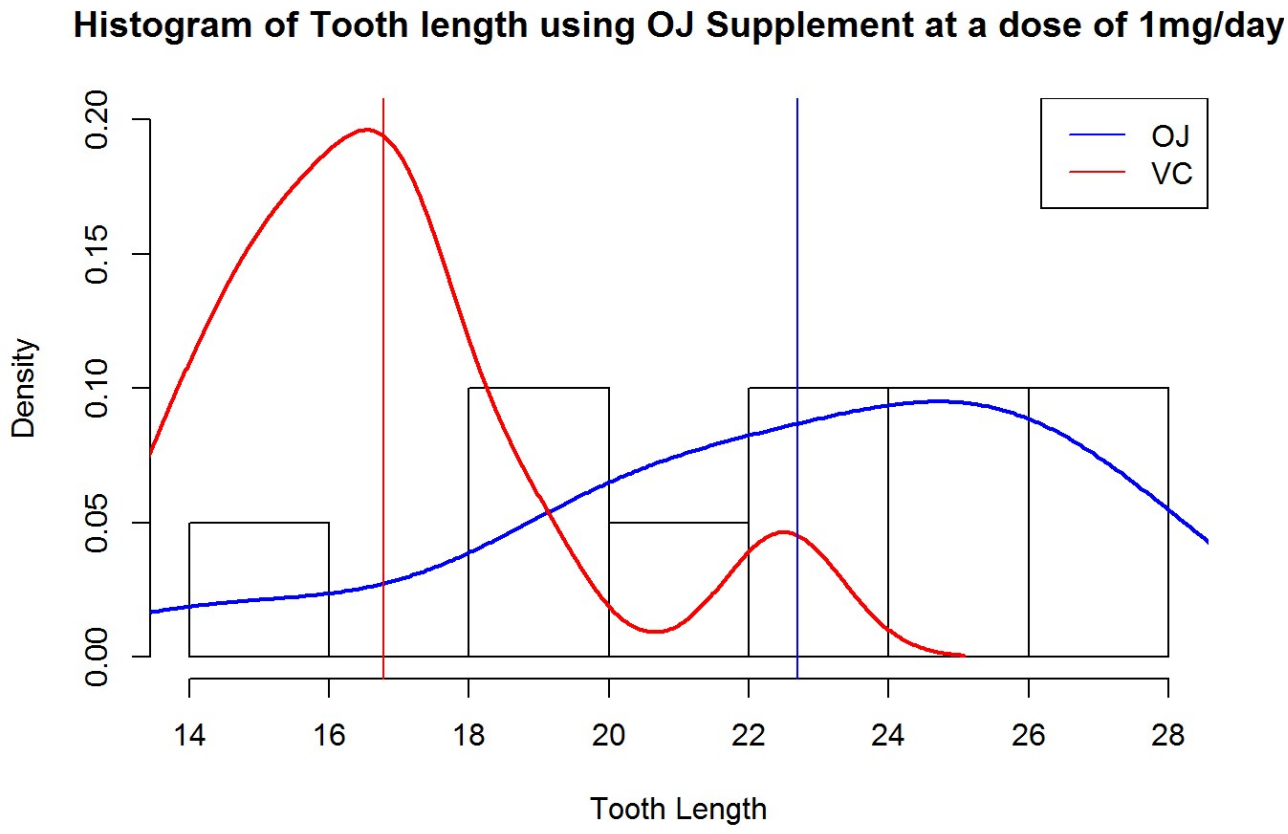
We will do the same as with the previous dosage and see if there is any significant difference in using orange juice versus vitamin c.

```
toothDose1 <- subset(ToothGrowth, dose == 1)
dose1test <- t.test(len ~ supp, paired = F, var.equal=F, data = toothDose1)
print(dose1test)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean in group OJ mean in group VC
##            22.70            16.77
```

Comparing the two supplements at this dosage there is quite a bigger difference compared to the 2 mg/day. The t-statistic is 4.0327696 which is larger and could be ground for rejecting the null hypothesis. The p-value (0.0010384) is now really small meaning there is a really low probably under the null hypothesis that we would see the t-statistic. This means either the null hypothesis is true and we have seen very unlikely means or we should reject the null hypothesis.

```
toothDose1_vc = toothDose1[toothDose1$supp == 'VC',]
toothDose1_oj = toothDose1[toothDose1$supp == 'OJ',]
hist(toothDose1_oj$len, probability = T, ylim=c(0,0.2),
     main='Histogram of Tooth length using OJ Supplement at a dose of 1mg/da
y', xlab = "Tooth Length")
lines(density(toothDose1_oj$len), col='blue', lwd=2)
lines(density(toothDose1_vc$len), col='red', lwd=2)
abline(v = dose1test$estimate[1], col='blue')
abline(v = dose1test$estimate[2], col='red')
legend('topright', c('OJ', 'VC'), lty = 1, col=c('blue', 'red'))
```



Dosage of 0.5 mg/day

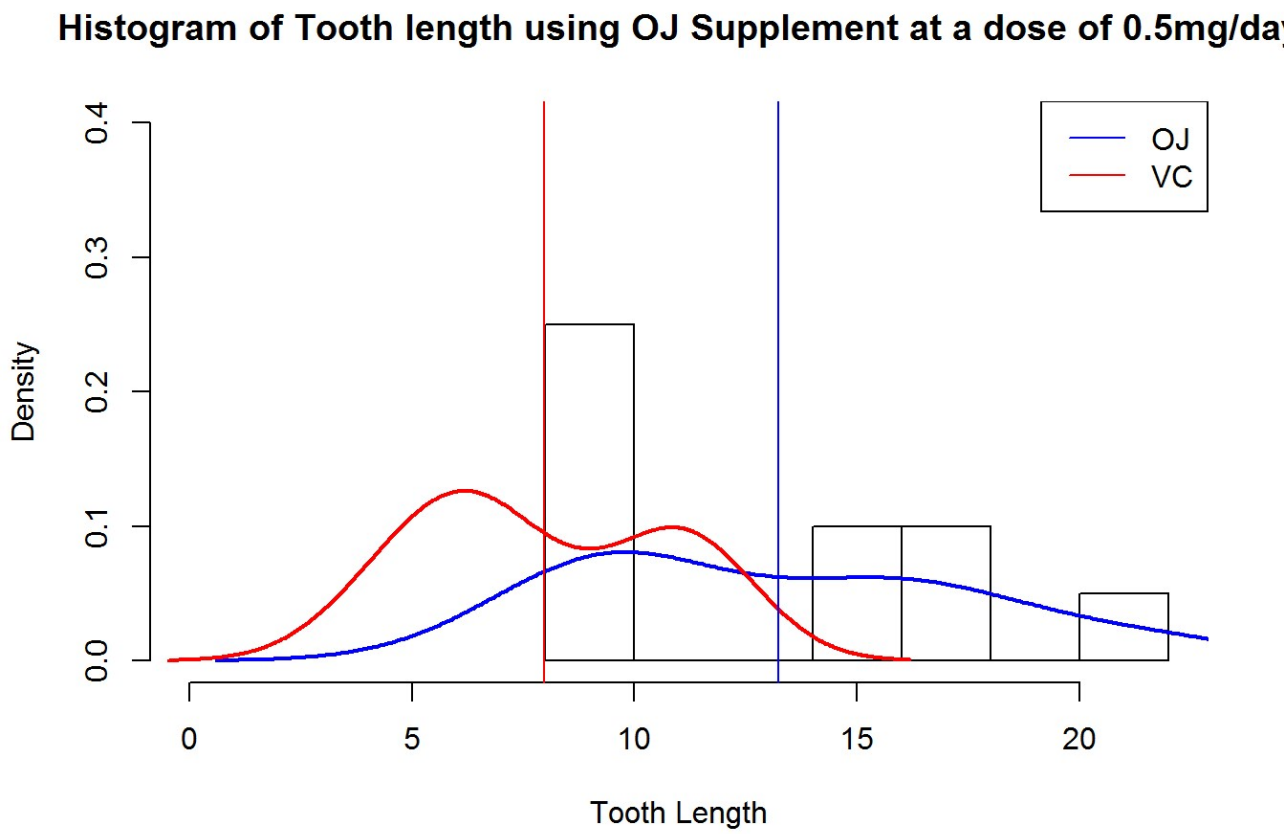
Now we do the same test for the last dosage of 0.5 mg/day.

```
toothDose05 <- subset(ToothGrowth, dose == 0.5)
dose05test <- t.test(len ~ supp, paired = F, var.equal=F, data = toothDose05)
print(dose05test)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
```

This comparison is very similar to the 1 mg/day dosage. In that the t-statistic(3.1697328) is higher and the p-value is low (0.0063586). Again we can confirm this visually by plotting the distributions.

```
toothDose05_vc = toothDose05[toothDose05$supp == 'VC',]
toothDose05_oj = toothDose05[toothDose05$supp == 'OJ',]
hist(toothDose05_oj$len, probability = T, ylim=c(0,0.4), xlim = c(0,22),
     main='Histogram of Tooth length using OJ Supplement at a dose of 0.5mg/da
y', xlab = "Tooth Length")
lines(density(toothDose05_oj$len), col='blue', lwd=2)
lines(density(toothDose05_vc$len), col='red', lwd=2)
abline(v = dose05test$estimate[1], col='blue')
abline(v = dose05test$estimate[2], col='red')
legend('topright', c('OJ', 'VC'), lty = 1, col=c('blue', 'red'))
```



Conclusion and Assumptions

The results of comparing the two supplements at equivalent dosages shows that at the higher dose of 2 mg/day there is no significant difference between orange juice and vitamin c. At doses of 1 mg/day and 0.5 mg/day there is a significant difference between the two supplements and you should chose one depending on your desired outcome. We can make these conclusions due to hypothesis testing and looking at the t-statistics, intervals and p-values. Some assumptions that were made were that this is not a paired test. Meaning the two supplement groups were unrelated. I also assumed that two groups did not have equal variance. We also made assumptions that let us use the t-test such as that our distribution is not skewed, roughly symmetric and mound shaped.